CLAIMS

What is claimed is:

1. An apparatus for clustering proteomic and genomic data, the apparatus comprising a computer system including a processor, a memory coupled with the processor, an input coupled with the processor for receiving proteomic and genomic data and for receiving user input, and an output coupled with the processor for outputting the clustered proteomic and genomic data, wherein the computer system further comprises means, residing in its processor and memory, for:

    a. receiving a set of data including *n* data samples, with each data sample having *m* characteristics;

    b. producing a one-dimensional ordering of the data samples, resulting in a linearly ordered set of data samples including *n-1* possible split points;

    c. configuring a dendrogram from the linearly ordered set of data samples by iteratively splitting the linearly ordered set of data samples into successive subsets and representing each split in the dendrogram until each subset contains one data sample by traversing the linearly ordered set of data samples and assigning a numerical quality value to each of the *n-1* possible split points with at least one of the numerical quality values being a best numerical quality value, and then splitting the set of data at at least one split point based on the best numerical quality values; and

    d. outputting the one-dimensional ordering of the data samples and the configuration of the dendrogram; whereby the data samples are clustered in order to allow for efficient analysis to be performed thereon.

2. An apparatus for clustering proteomic and genomic data, as set forth in claim 1, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality

value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

3. An apparatus for clustering proteomic and genomic data, as set forth in claim 1, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

4. An apparatus for clustering proteomic and genomic data, as set forth in claim 1, wherein the means for producing the one-dimensional ordering of the data samples is principal component analysis.

5. An apparatus for clustering proteomic and genomic data, as set forth in claim 4, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

6. An apparatus for clustering proteomic and genomic data, as set forth in claim 4, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

7. An apparatus for clustering proteomic and genomic data, as set forth in claim 1, wherein the means for producing the one-dimensional ordering of the data samples is a one-dimensional, self-organizing map.

8. An apparatus for clustering proteomic and genomic data, as set forth in claim 7, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

9. An apparatus for clustering proteomic and genomic data, as set forth in claim 8, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each

possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

10. An apparatus for clustering proteomic and genomic data, the apparatus comprising a computer system including a processor, a memory coupled with the processor, an input coupled with the processor for receiving proteomic and genomic data and for receiving user input, and an output coupled with the processor for outputting the clustered proteomic and genomic data, wherein the computer system further comprises, residing in its processor and memory:

a. a receiving module for receiving a set of data including *n* data samples, with each data sample having *m* characteristics;

b. an ordering module for producing a one-dimensional ordering of the data samples, resulting in a linearly ordered set of data samples including *n-1* possible split points;

c. a dendrogram module for configuring a dendrogram from the linearly ordered set of data samples by iteratively splitting the linearly ordered set of data samples into successive subsets and representing each split in the dendrogram until each subset contains one data sample by traversing the linearly ordered set of data samples and assigning a numerical quality value to each of the *n-1* possible split points with at least one of the numerical quality values being a best numerical quality value, and then splitting the set of data at at least one split point based on the best numerical quality values; and

d. an output module for outputting the one-dimensional ordering of the data samples and the configuration of the dendrogram; whereby the data samples are clustered in order to allow for efficient analysis to be performed thereon.

11. An apparatus for clustering proteomic and genomic data, as set forth in claim 10, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

12. An apparatus for clustering proteomic and genomic data, as set forth in claim 10, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

13. An apparatus for clustering proteomic and genomic data, as set forth in claim 10, wherein the ordering module is principal component analysis module.

14. An apparatus for clustering proteomic and genomic data, as set forth in claim 13, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is

representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

15. An apparatus for clustering proteomic and genomic data, as set forth in claim 13, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

16. An apparatus for clustering proteomic and genomic data, as set forth in claim 10, wherein the ordering module is a one-dimensional, self-organizing map.

17. An apparatus for clustering proteomic and genomic data, as set forth in claim 16, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

18. An apparatus for clustering proteomic and genomic data, as set forth in claim 16, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

19. A method for clustering proteomic and genomic data on a computer system including a processor, a memory coupled with the processor, an input coupled with the processor for receiving proteomic and genomic data and for receiving user input, and an output coupled with the processor for outputting the clustered proteomic and genomic data, wherein the method comprises the steps of:

a. receiving a set of data including *n* data samples, with each data sample having *m* characteristics;

b. producing a one-dimensional ordering of the data samples, resulting in a linearly ordered set of data samples including *n-1* possible split points;

c. configuring a dendrogram from the linearly ordered set of data samples by iteratively splitting the linearly ordered set of data samples into successive subsets and representing each split in the dendrogram until each subset contains one data sample by traversing the linearly ordered set of data samples and assigning a numerical quality value to each of the *n-1* possible split points with at least one of the numerical quality values being a best numerical quality value, and then splitting the set of data at at least one split point based on the best numerical quality values; and

d. outputting the one-dimensional ordering of the data samples and the configuration of the dendrogram; whereby the data samples are clustered in order to allow for efficient analysis to be performed thereon.

20. A method for clustering proteomic and genomic data, as set forth in claim 19, wherein the step of configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

21. A method for clustering proteomic and genomic data, as set forth in claim 19, wherein the step of configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

22. A method for clustering proteomic and genomic data, as set forth in claim 19, wherein step of producing the one-dimensional ordering of the data samples is performed by principal component analysis.

23. A method for clustering proteomic and genomic data, as set forth in claim 22, wherein the step of configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split

point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

24. A method for clustering proteomic and genomic data, as set forth in claim 22, wherein the step of configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

25. A method for clustering proteomic and genomic data, as set forth in claim 19, wherein the step of producing the one-dimensional ordering of the data samples is performed by a one-dimensional, self-organizing map.

26. A method for clustering proteomic and genomic data, as set forth in claim 25, wherein the step of configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each

successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

27. A method for clustering proteomic and genomic data, as set forth in claim 25, wherein the step of configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

28. A computer program product for clustering proteomic and genomic data, the computer program product comprising means, stored on a computer readable medium, for:

a. receiving a set of data including *n* data samples, with each data sample having *m* characteristics;
b. producing a one-dimensional ordering of the data samples, resulting in a linearly ordered set of data samples including *n-1* possible split points;
c. configuring a dendrogram from the linearly ordered set of data samples by iteratively splitting the linearly ordered set of data samples into successive subsets and representing each split in the dendrogram until each subset contains one data sample by traversing the linearly ordered set of data samples and assigning a numerical quality value to each of the *n-1* possible split points with at least one of the numerical quality values being a best numerical quality value, and then splitting the set of data at at least one split point based on the best numerical quality values; and

d. outputting the one-dimensional ordering of the data samples and the configuration of the dendrogram; whereby the data samples are clustered in order to allow for efficient analysis to be performed thereon.

29. A computer program product for clustering proteomic and genomic data, as set forth in claim 28, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

30. A computer program product for clustering proteomic and genomic data, as set forth in claim 28, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

31. A computer program product for clustering proteomic and genomic data, as set forth in claim 28, wherein the means for producing the one-dimensional ordering of the data samples is principal component analysis.

32. A computer program product for clustering proteomic and genomic data, as set forth in claim 31, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

33. A computer program product for clustering proteomic and genomic data, as set forth in claim 31, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

34. A computer program product for clustering proteomic and genomic data, as set forth in claim 28, wherein the means for producing the one-dimensional ordering of the data samples is a one-dimensional, self-organizing map.

35. A computer program product for clustering proteomic and genomic data, as set forth in claim 34, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the

numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

36. A computer program product for clustering proteomic and genomic data, as set forth in claim 34, wherein the means for configuring the dendrogram iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

37. A computer program product for clustering proteomic and genomic data, the computer program product stored on a computer readable medium, and comprising:

a. a receiving module for receiving a set of data including *n* data samples, with each data sample having *m* characteristics;

b. an ordering module for producing a one-dimensional ordering of the data samples, resulting in a linearly ordered set of data samples including *n-1* possible split points;

c. a dendrogram module for configuring a dendrogram from the linearly ordered set of data samples by iteratively splitting the linearly ordered set of data samples into successive subsets and representing each split in the dendrogram until each subset contains one data sample by traversing the linearly ordered set of data samples and assigning a numerical quality value to each of the *n-1* possible split points with at least one of the

numerical quality values being a best numerical quality value, and then splitting the set of data at at least one split point based on the best numerical quality values; and

d. an output module for outputting the one-dimensional ordering of the data samples and the configuration of the dendrogram; whereby the data samples are clustered in order to allow for efficient analysis to be performed thereon.

38. A computer program product for clustering proteomic and genomic data, as set forth in claim 37, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

39. A computer program product for clustering proteomic and genomic data, as set forth in claim 37, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

40. A computer program product for clustering proteomic and genomic data, as set forth in claim 37, wherein the ordering module is principal component analysis module.

41. A computer program product for clustering proteomic and genomic data, as set forth in claim 40, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

42. A computer program product for clustering proteomic and genomic data, as set forth in claim 40, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.

43. A computer program product for clustering proteomic and genomic data, as set forth in claim 37, wherein the ordering module is a one-dimensional, self-organizing map.

44. A computer program product for clustering proteomic and genomic data, as set forth in claim 43, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a local quality technique, in which a numerical quality value is assigned to each possible split point, where each split point resides between two adjacent data samples and where the numerical quality value for each split point is representative of the distance between the two adjacent data samples between which the split point resides, with the data set being split at the split point having the greatest quality value, so that each successive split of the data set provides two data subsets with each of the subsets including the data samples on a respective side of the split point.

45. A computer program product for clustering proteomic and genomic data, as set forth in claim 43, wherein the dendrogram module iteratively splits the linearly ordered set of data samples by using a within-group variance technique, in which a numerical quality value is assigned to each possible split point, where at each possible split point the set of data samples is divided into two sides, with the numerical quality value at each possible split point being the sum of the variances of the data samples on each side of the split point, and where the splitting of the set of data samples occurs at the split point with the lowest such within-group variance, resulting in two linearly-ordered data sample subsets.